Entity-based Sentiment Classifier for Social Media Analysis

Master Thesis Seminar

Presented by
Cristobal Leiva
Supervisor
Dr. Simon Scerri

Evaluators
Prof. Dr. Sören Auer
Prof. Dr. Jens Lehmann

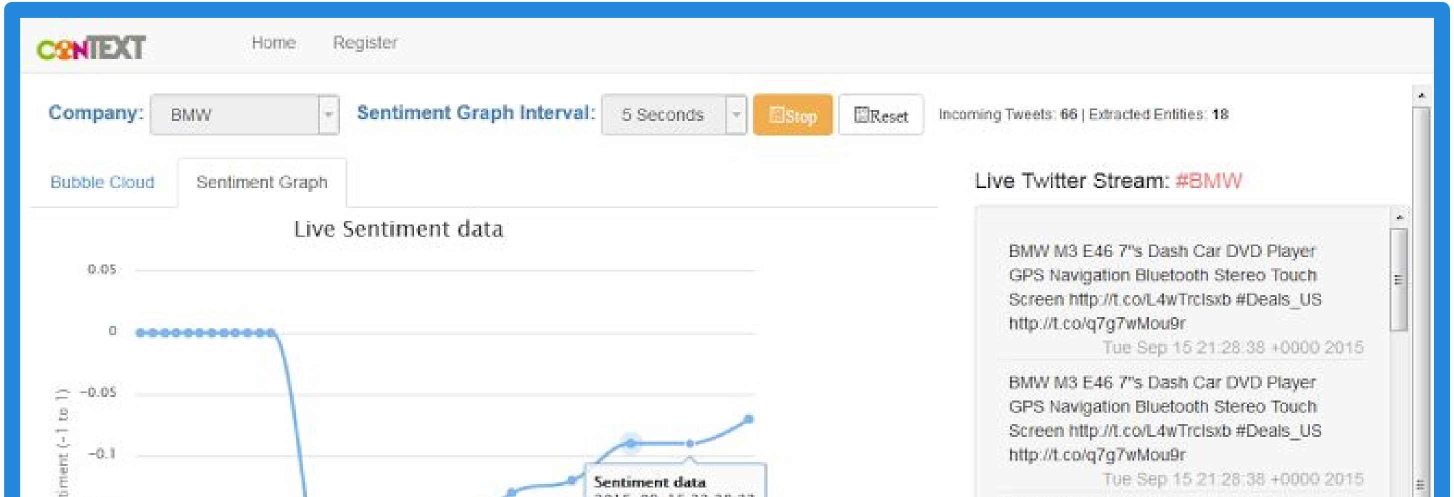
universitätbonn
EIS
ENTERPRISE INFORMATION SYSTEMS

Motivation

- Social media networking services such as **Twitter** provide a massive amount of valuable data.
- Core business processes such as **market-sensing**, customer acquisition and customer relationship management (CRM).
- Cross domain applications: **Politics**, Sociology and others.

Motivation - SentiTrack

- **Linked Data-based Social Media Analysis for Stock Market Tracking**
    - **ReSA** (Real-time Sentiment Analysis) by Dr. Ali Khalili
    - Find correlation between **public sentiments** and intra-day **stock prices**

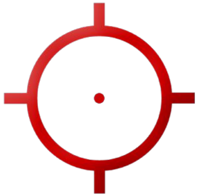


Problem definition

- **State-of-the-art approaches**
  - **Entities** are usually ignored
  - Only **document-level** analysis
  - Not designed for **tweets**
  - Presence of **multiple** entities
- **Real-Time Systems**
  - **Performance** issues

Objective

***"Classify* the sentiment of tweets according to the opinion expressed towards a *target entity*"**



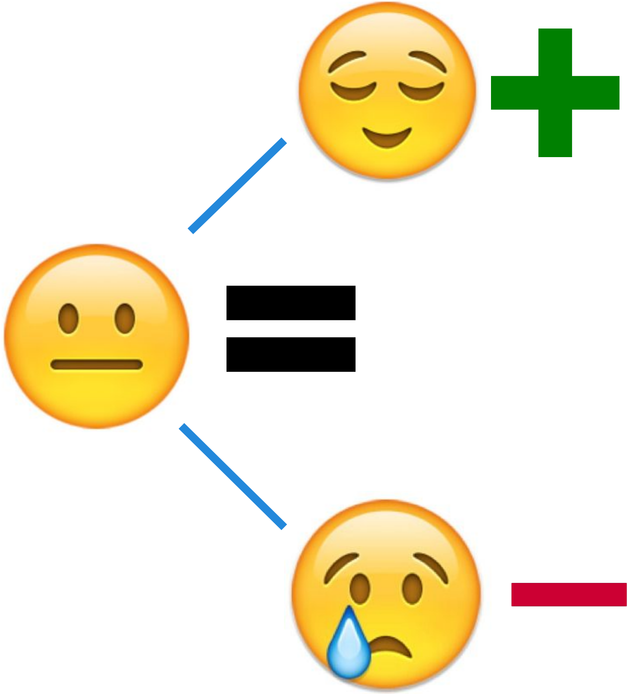
"my iPhone is <u>better than</u> your Nexus 4"

Entity *Entity*

Solution Proposed

- **Entity-based Sentiment Classifier**
  - **3-Class** Machine Learning approach
  - **Identify** and categorize entities
  - Entity **context** extraction
  - Entity-based **features**
- **Datasets for training**
  - Collection of **target-labeled** tweets

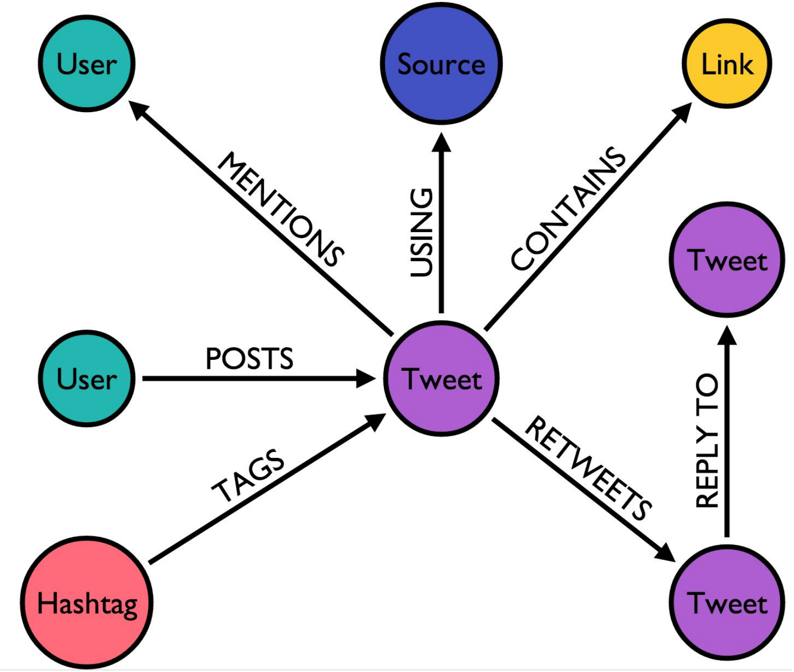
Background - Sentiment Analysis



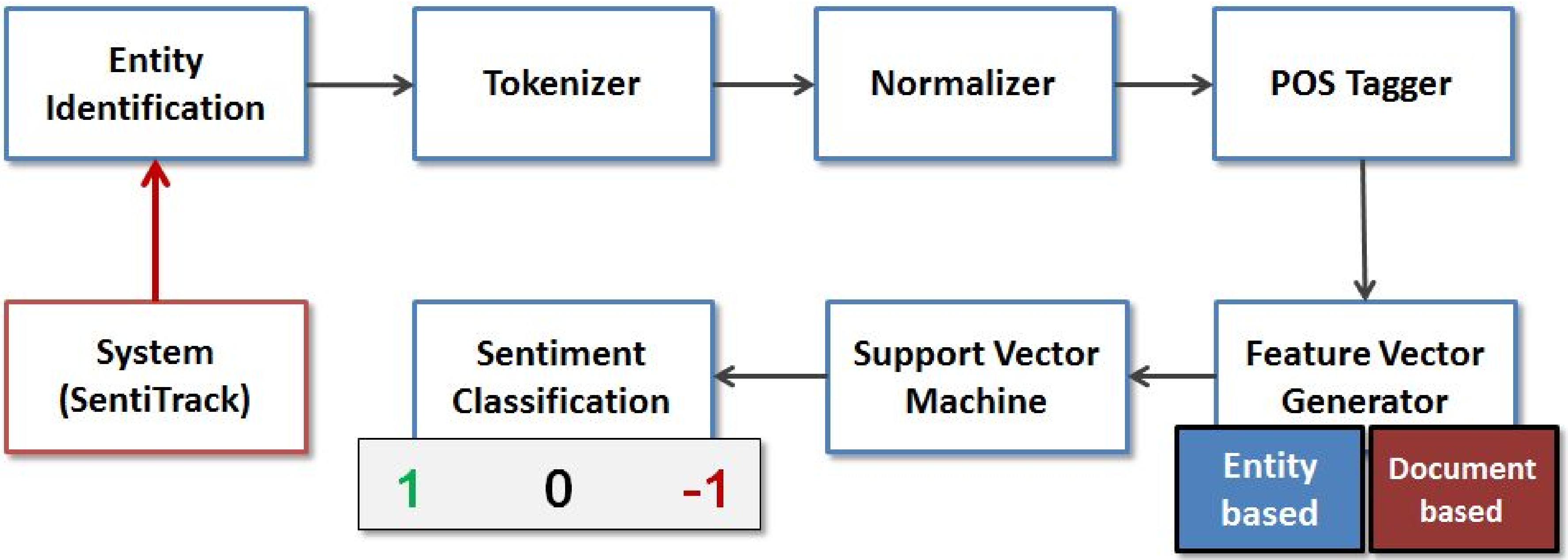
- **Extracting Opinions**
  - NLP / IR Task
  - Analysis **levels** (3)
- **Classifiers**
  - Supervised / Unsupervised
  - Classification **classes**
  - Polarity / Subjectivity

Background - Twitter

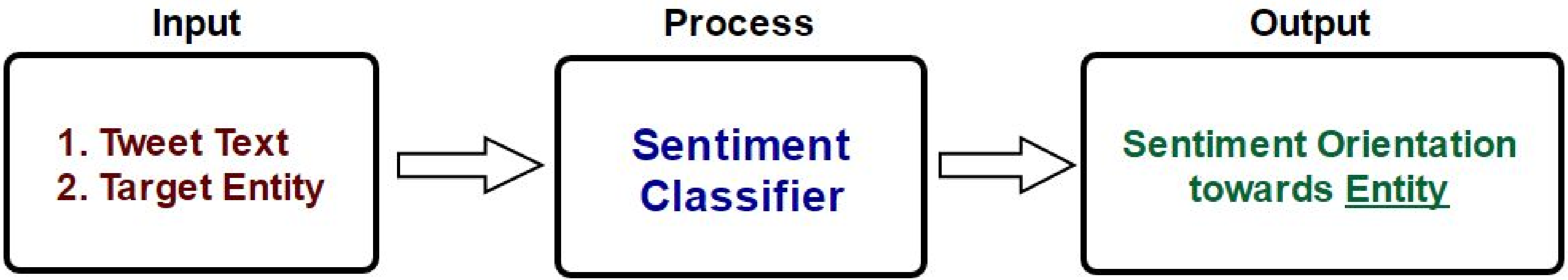
- **Features**
  - Short and concise
  - Slang / Abbr.
  - Unique **elements**
- **Twitter API**
  - **Real-Time** analysis
  - **Query** based search



Approach - Overview

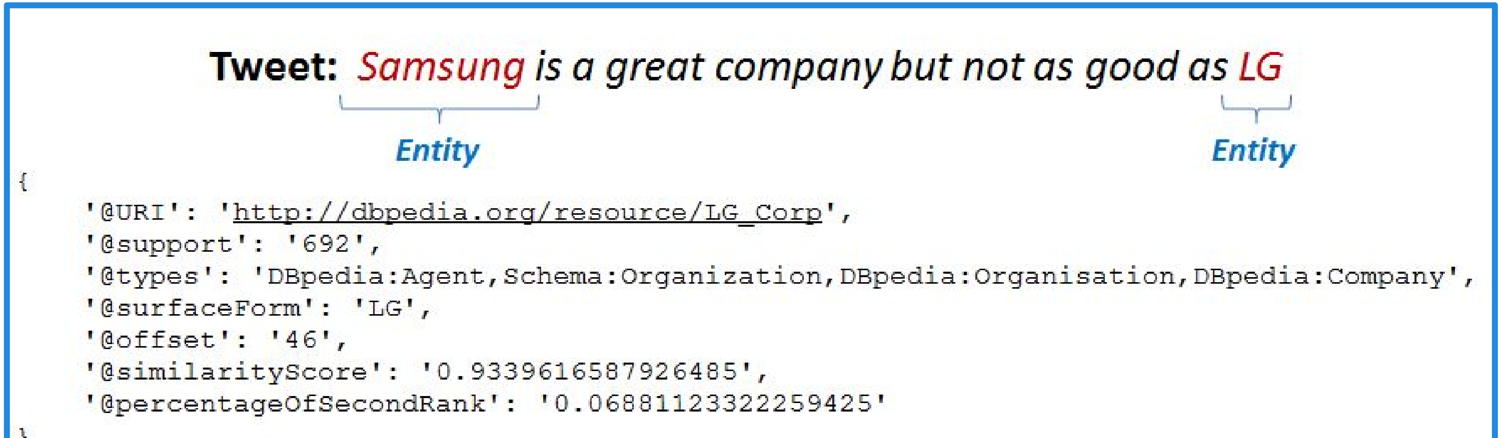


Approach - Workflow



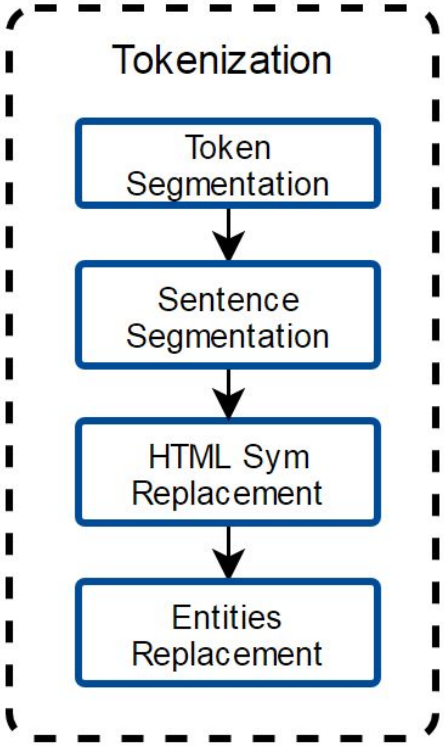
Approach - Entity Identification

- **DBpedia Spotlight**
  - Input **labeled** entities (Company / Person / Product) from SentiTrack
  - **Forward array** of entities (Target and Others) to Tokenizer module

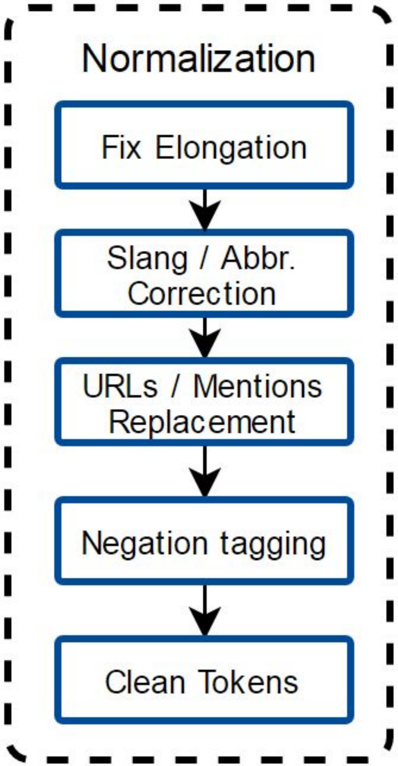


Approach - Tokenizer

- **Tokenization process**
  - **Splitting** of tweets into tokens using Regex
  - Identify sentence finalization tokens to for sentence separation (Regex)
  - Replace **HTML elements** with real values
  - Entities are replaced by **placeholders** to reduce sparsity of feature vectors.



Approach - Normalizer



- **Normalization process**
  - Fix tokens with more than two **repeated** letters
  - Replace slang and abbreviations with **correct form** (Urban Dictionary)
  - Replace URLs and @Mentions with **placeholders**
  - Tag tokens with "_NEG" that follow a **negated word** (not, never, none, etc...)
  - Remove numbers and symbols (Except Emojis)

Approach - Example Normalizer / Tokenizer

Target Entity	(1) Google
Other Entities	(1) Nexus
Tweet	Thanks google!! Just got my new Nexus <3
Result Tokens	(1) {Thanks, TargetEntity!!} (2) {Just, got, my, new, OtherEntity, <3}

Tokenized

Normalized

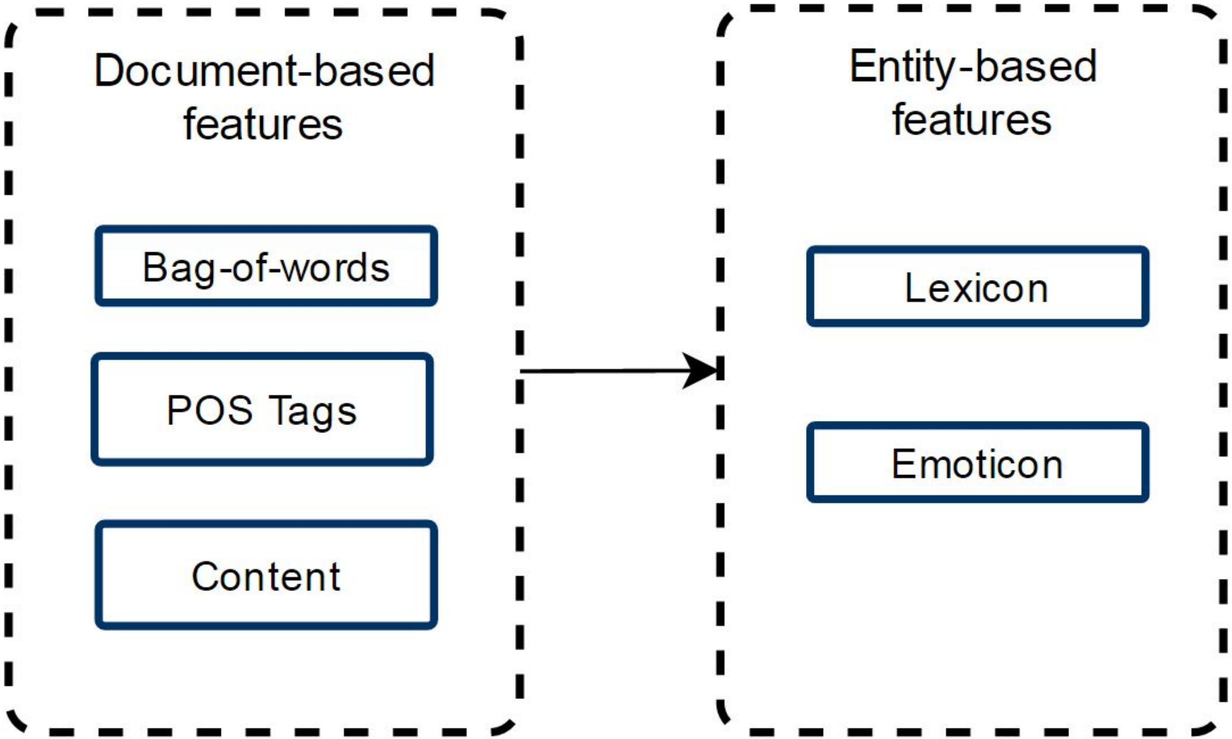
Tokens	**Normalization Result**
(1){not, their, best, !}	(1){not, their_NEG, best_NEG, !}
(2){ http://t.co, @Muse, #LiveMuse}	(2){ someURL, someUser, #LiveMuse}

Approach - POS Tagger

- **Part-of-speech tags on tokens**
  - Twitter ARK POS Tagger (Carnegie Mellon Uni) java-based
  - Most relevant tags are: *Nouns* / *Adjectives* / *Verbs* / *Adverbs*

Normalized Tokens	POS Tagging
(1){not, their_NEG, best_NEG}	{R/not, O/their_NEG, A/best_NEG}
(2){ someURL, someUser, #Live}	{ someURL, someUser, #/#Live}

Approach - Feature Vector Generator



FVG / Document - Binary Bag-of-words

- **Boolean term frequency**
  - Vector space reduced by removal of Stopwords
  - Values can only be 1 or 0, present or not

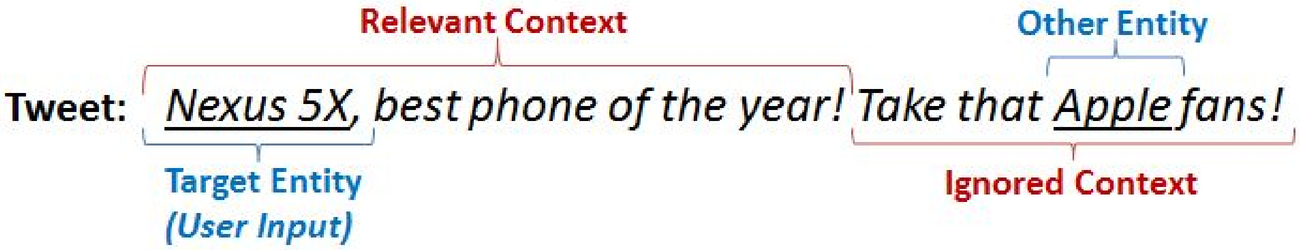
Tweets	Binary Bag-of-words
happy birthday friend! :)	{1,1,1,1,0,0,0}
always be happy ;)	{1,0,0,0,1,1,1}

FVG / Document - POS Tags / Content

- **POS Tags Features**
  - No. of Nouns / Adjectives / Verbs / Adverbs
- **Content Features**
  - No. of all-caps, hashtags, elongated words, negation context, punctuation.

FVG / Entity - Context Extraction

- **Sentence separation**
  - Ignore **non-target sentences** (Outside Neighborhood) for Entity features
- **“But” clause and conditions**
  - Rules for condition expression (“except that, but, better than”)



FVG / Entity - Lexicon

Lexicons	Score Range	Words
MaxDiff	Real-values	1,500
AFINN	-5 to 5	2,477
BingLiu	Pos / Neg	6,785
SentiWordNet	-1 to 1	147,292
MPQA	Pos / Neg	6,886
NRC Hashtag	Real-values	54,129
Sentiment140	Real-values	62,468

- **Lexical Resources**
  - Different **score ranges**
  - All 3 types of lexicons
- **Calculations**
  - **No.** sentiment tokens
  - **Total** score
  - **Max** score
  - **Last** token score

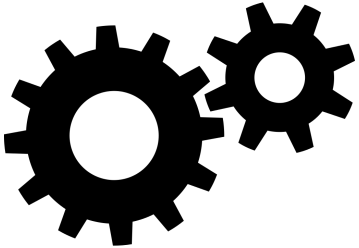
FVG / Entity - Emoticons / Example

- **Emoticon Features**
  - Resource Pos / Neg emojis (SentiStrength)
  - No. of Pos / No. of Neg

Target-entity context tokens	Feature Vectors
{my, TargetEntity, is, awesome,	(BingLiu){2, 2, 1, 1}
best, day, ever, :D }	(Emoji){1,0}

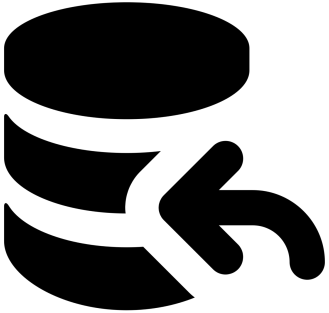
Approach - Support Vector Machine / SentiTrack

- **Node-SVM (LibSVM)**
  - Linear kernel with **default** parameters
  - No class weighting (balanced training data)
  - Active sparse format to **ignore** 0 in vector space
- **SentiTrack - common technologies**
  - **Node-js** implementations
  - Modular integration with **.npmjs** (Node Package Manager)

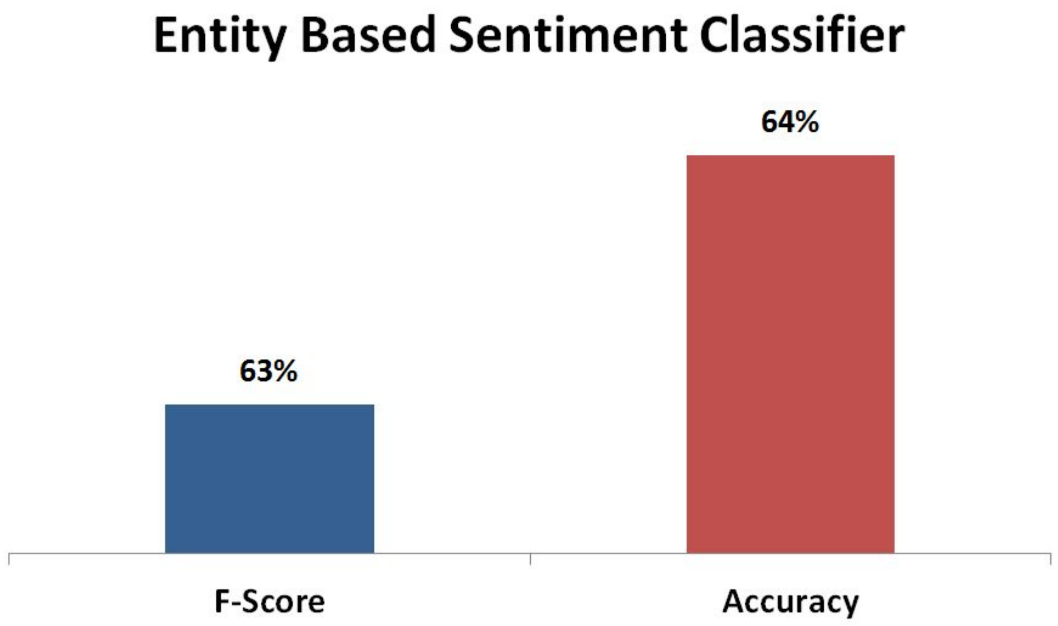


Evaluation Results - Datasets

- **Collection**
  - Semeval 2015 (**Semantic Evaluation**) - Task 10 - Training data
  - Semeval 2016 - Task 4 - Training data
  - Twitter **Sanders** Analytics Corpus
  - **STS** - Gold (Saif M. Mohammad)
- **Data Summary**
  - **4900** entity-based annotated tweets (balanced classes)
  - **70%** - 30% SVM Training / Eval ratio

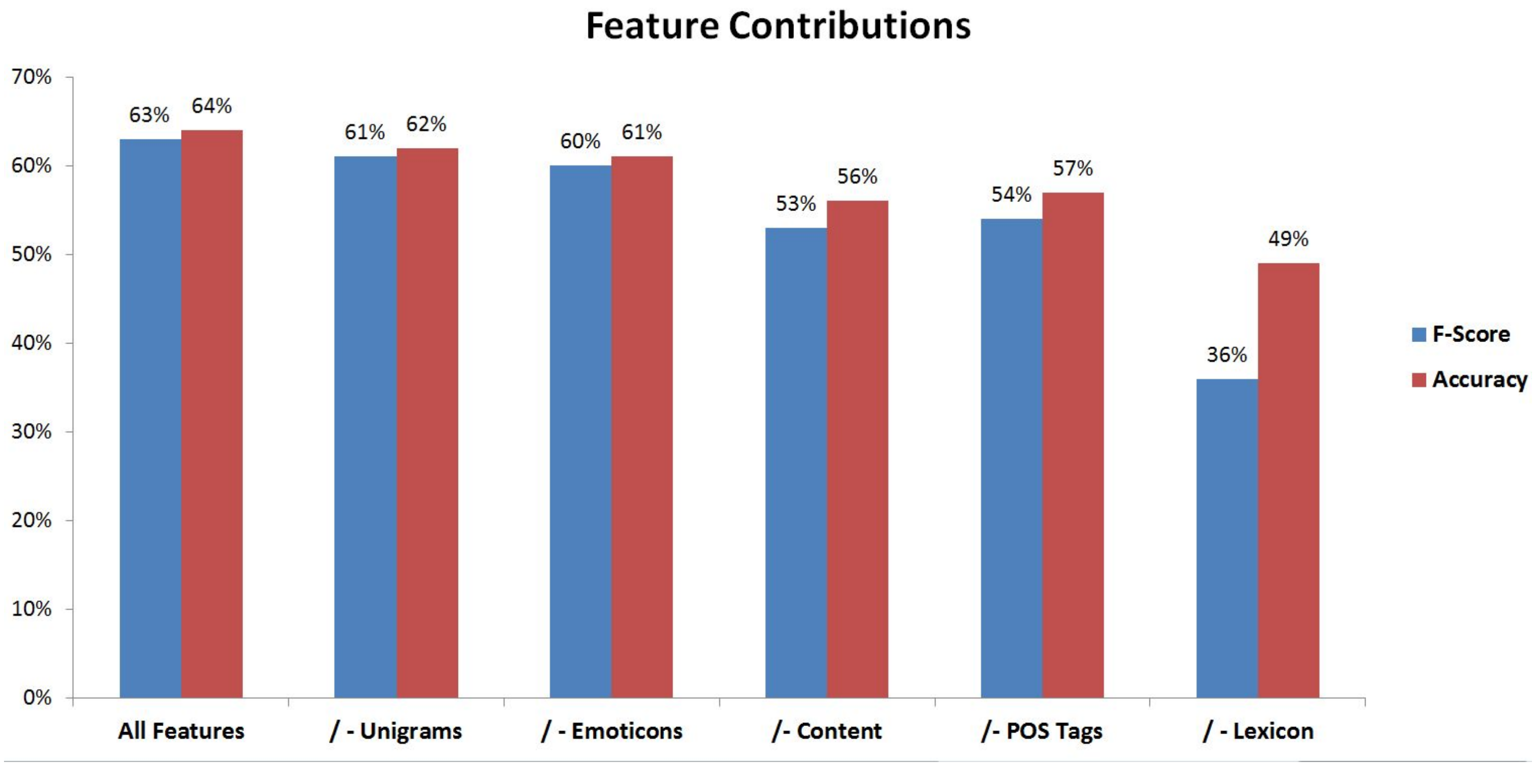


Evaluation Results - Quality

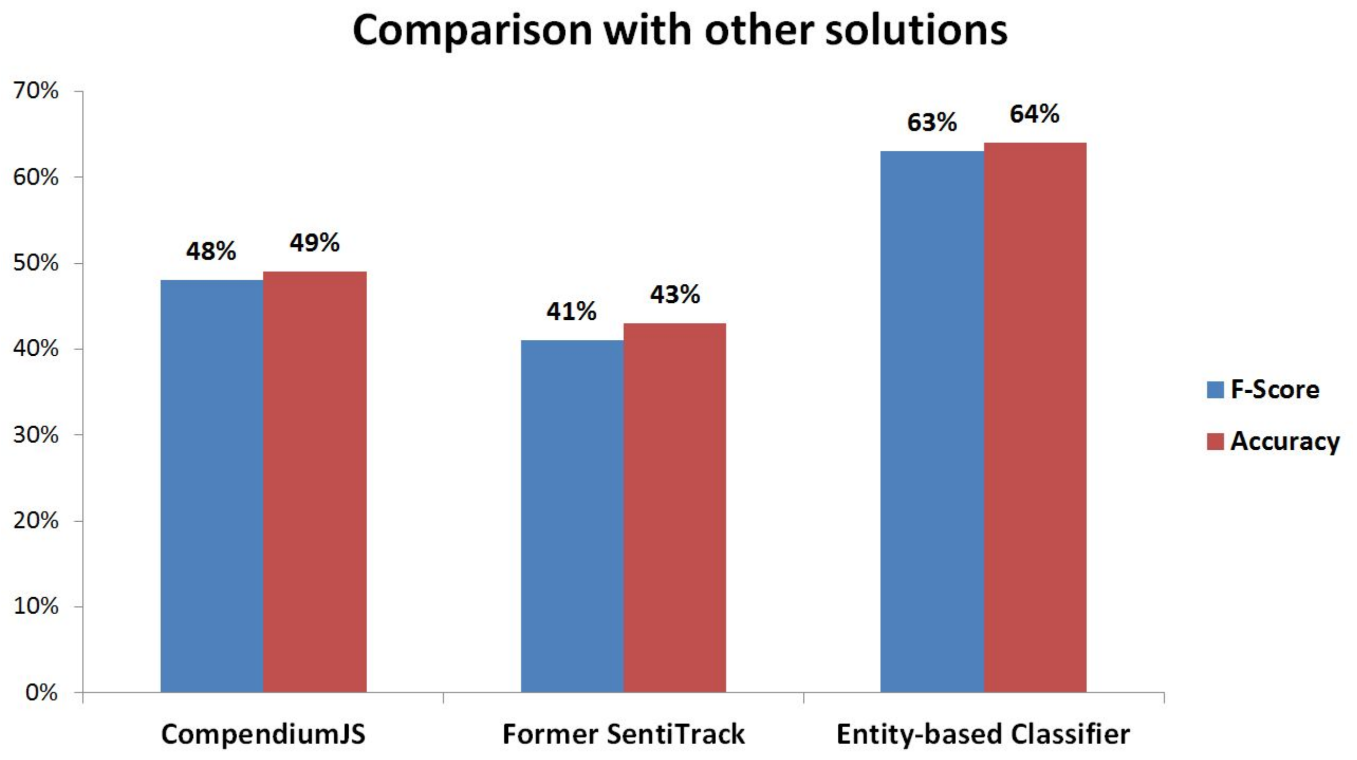


- **Evaluation Metrics**
  - F-Score / Accuracy
  - 4-fold **cross validation**
  - **Features** quality test
  - **70%** - 30% training / eval ratio

Evaluation Results - Quality



Evaluation Results - Comparison



Evaluation Results - Performance

- **Evaluation environment**
  - Intel Core i5-2320 CPU @ 3.00GHz
  - 8 GB RAM
  - 64 bits Windows 7
- **Results for 1000 tweets processed**
  - Former SentiTrack Classifier: **323 ms**
  - CompendiumJS: **2357 ms**
  - Entity-based Classifier: **3447 ms**

Conclusions

- This thesis presented the **research, solution and evaluation** of an entity-based sentiment classifier for social media analysis
- Implemented sentiment classifier was **fully integrated** to SentiTrack and ReSA. Which proves its compatibility with real-time systems
- Proposed approach achieved **satisfactory results** with 20% accuracy **improvement** over former SentiTrack classifier

Future Work

- Expand the sentiment classifier with a **dependency parser** module capable of performing real time analysis
- Improve the quality of the classifier by including higher level **n-grams** and a better **Named Entity Recognition** module
- Develop a **graphical user interface** that allows users to classify tweets
- Enhance Support Vector Machine module by the inclusion of **distant supervision** methods

SentiTrack...

Thank You

Presented by
Cristobal Leiva

Supervised by
Dr. Simon Scerri
Prof. Dr. Sören Auer
Prof. Dr. Jens Lehmann